DOCUMENT RESUME

ED 368 791 TM 021 310

AUTHOR Ito, Kyoko; Sykes, Robert C.
TITLE The Effect of Restricting Ability Distributions in the Estimation of Item Difficulties: Implications for a CAT Implementation.
PUB DATE Apr 94
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; *Computer Assisted Testing; *Difficulty Level; Estimation (Mathematics); Item Banks; Item Response Theory; Licensing Examinations (Professions); Reference Groups; Sample Size; Simulation; Statistical Distributions; *Test Items
IDENTIFIERS B Values; Calibration; Paper and Pencil Tests; Rasch Model; *Recalibration

ABSTRACT

Responses to previously calibrated items administered in a computerized adaptive testing (CAT) mode may be used to recalibrate the items. This live-data simulation study investigated the possibility, and limitations, of on-line adaptive recalibration of precalibrated items. Responses to items of a Rasch-based paper-and-pencil licensure examination were used to simulate CAT and paper-and-pencil administrations, defining CAT forms of varying difficulty levels and samples of various sizes. Forms were calibrated and new b-values were compared with the bank b-values obtained from responses of the reference group taking the paper-and-pencil examination. Results indicate that bank b-values were not well replicated when difficult items were calibrated using responses from able examinees and easy items were calibrated using responses from less able examinees. On-line adaptive recalibration as simulated in this study has limitations. However, a "modified" on-line adaptive recalibration may still be a possibility as long as: (1) a reasonably large CAT recalibration sample is predefined from the reference group; (2) the sample has a mean ability similar to that of the reference group; and (3) items to be recalibrated together are relatively heterogeneous in Rasch difficulty. Eight tables are included. (Contains 3 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 368 791

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kyoko Ito

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Effect of Restricting Ability Distributions in the Estimation of Item Difficulties: Implications for a CAT Implementation

Kyoko Ito
Robert C. Sykes

CTB McGraw-Hill

This paper was presented in April, 1994 at the Annual Meeting of the National Council on Measurement in Education
New Orleans, LA

TM021310

BEST COPY AVAILABLE

Introduction

Computerized adaptive tests (CAT) cannot be successfully administered without reliable item parameter estimates. In the first few years of operation, examinees can be scored using item parameter estimates obtained from previous paper-and-pencil administrations of the items, provided no mode effect exists. Under these circumstances, field-test items may also be calibrated "on-line" from administrations of linear computerized tests or linear portions of CAT (i.e., **on-line non-adaptive calibration**). Because these tests are not targeted on examinee ability, the resulting parameter estimates from linear computerized tests are expected to be similar to those obtained from paper-and-pencil random samples.

Previously scored items will likely have to be recalibrated after a period of operational administrations, for reasons such as the need to evaluate scale drift. An obvious way to recalibrate the items would be on-line non-adaptive recalibration. In this case, previously scored items will be treated just like field-test items; namely, items will not be selected based on their parameter estimates, but rather, they will be (quasi-)randomly administered to examinees. Because items are administered to examinees having a wide range of ability, on-line non-adaptive recalibration of precalibrated items should yield parameter estimates similar to those from a paper-and-pencil calibration.

A drawback of the on-line non-adaptive recalibration of precalibrated items is that such non-CAT, random administrations of items add costs to normal CAT operations. Time and cost savings from eliminating random administrations of items could be considerable if previously scored items could be recalibrated using responses to the items administered adaptively in operational CAT sessions. This procedure may be called **on-line adaptive recalibration**, as opposed to on-line non-adaptive recalibration.

Before on-line adaptive recalibration may be considered a viable option, however, the effects on item parameter estimates of targeting items to the examinee's ability need to be investigated. Using responses to a one-parameter (Rasch) based paper-and-pencil examination, a "live-data" simulation was conducted to evaluate the effects of targeting items on item parameter estimates or b-values.

Source Data

The source data were examinee responses to the 298 scored items in a form of a paper-and-pencil (hereafter abbreviated as "PP") licensure examination of a health-care related profession.

All scored items in the full-length form had been screened for the fit of reference-group responses to the Rasch model prior to their assignment to the form. Once items were selected for simulated forms (see below), responses to the items in each form were extracted from the source data and analyzed.

Benchmark Bank B-Values and the Reference Group of Examinees

The item parameters of concern were one-parameter b-values. B-values obtained from simulated forms were compared with bank b-values. Bank b-values had been obtained by, first, calibrating the PP items using responses from a sample of 1,000 examinees from the reference group and then equating them to the bank scale so that they would have the same mean as the mean of the b-values obtained from their last previous paper-and-pencil administration. These bank b-values were used throughout this study as benchmark values. PARMATE (Burket, 1990), a program written at CTB Macmillan/McGraw-Hill, was used in all item calibrations in the study. PARMATE provides maximum likelihood estimates using the procedure described by Lord (1980, pp.180-181).

The reference group for this examination consisted of examinees who received an education in the United States and took the examination (PP) for the first time ("First-Time U.S."). The mean ability estimate for the reference group was 0.6 logits for the full-length PP form.

Study Design

Items (or forms) and examinees (or samples) were selected to simulate a CAT administration and a paper-and-pencil administration. For the CAT simulations, **three "CAT forms"** and **three "CAT samples"** were defined. The CAT forms were constructed to contain items from three separate spans of item difficulty -- **Hard, Medium Difficulty, and Easy.** In CAT sessions, hard items would tend to be seen only by very able examinees, and easy items by less competent examinees. Based on this logic, within each of the subpopulations of interest, CAT samples of varying sizes were selected from the three ranges of ability that roughly corresponded to the three ranges of item difficulty. They are referred to as **High, Middle, and Low Ability.**

For the paper-and-pencil simulations, **two shorter versions of the PP -- "Mini PP" and "Shorter PP"** -- were constructed. To simulate conventional paper-and-pencil samples, **"Representative Samples"** of various sizes were drawn randomly from each of the subpopulations. Paper-and-pencil samples would be representative of the population in the absence of targeting of items to

ability.

All forms conformed to the test-plan specifications.

Characteristics of the Forms

Table 1 presents descriptive statistics for the forms. The middle four columns in the table ("Bank B-Value") give the information on the forms' bank b-values. The two right columns in the table, "Abs. 1st Lds." and "Abs. 2nd Lds," are the means and standard deviations of the absolute values of the first and second factor loadings from the Stout assessments of essential dimensionality of the forms (Stout, 1987, 1990). Examination forms have consistently been demonstrated to be bidimensional by the Stout criterion, although the bidimensionality has been shown to have no practical effects on pass/fail classification rates.

As shown in Table 1, the CAT forms were comparable in length. The Hard and Medium Difficulty CAT forms contained 73 items each, while the Easy CAT form contained slightly fewer items (68 items). Although these CAT forms were much shorter than the PP, they were of a comparable length to the smallest length CAT tests that will be administered in the client's CAT program.

The relative difficulty of the three CAT forms is evident in the lowest and highest b-values, as well as in the mean bank b-values. The bank b-values of the Hard CAT form ranged from -0.63 to 0.86, with a mean of -0.02 (logits). The items in the Medium Difficulty CAT form were less difficult, having b-values ranging from -1.17 to -0.55, with a mean of -0.82. The Easy CAT form contained items with bank b-values ranging between -1.96 to -0.98, with a mean of -1.34. Note that the CAT forms differed in the standard deviation of b-values. All three CAT forms had means and standard deviations of factor loadings similar to those of the PP.

Two shorter versions of the PP were the 75-item Mini PP and the 149-item Shorter PP. The Mini PP was thus comparable to the simulated CAT forms with regard to test length. Both the Mini PP and Shorter PP were similar to the CAT forms in terms of the mean absolute first and second factor loadings. For comparison purposes, Table 1 also shows the statistics for the full-length PP. It is clear from the table that the Mini PP and Shorter PP were substantially comparable to the PP except for test length.

Characteristics of the Samples

The CAT samples were selected using examinees' ability estimates based on the total of 298 scored items in the PP. The three ability bands from which the CAT samples were drawn are shown below, along with the corresponding b-value ranges.

Items : B-Value Range	CAT Samples: Theta Range
-0.63 - 0.86 (Hard)	-0.5 - 0.0 (High)
-1.17 - -0.55 (Medium)	-1.0 - -0.5 (Middle)
-1.96 - -0.98 (Easy)	-1.7 - -1.05 (Low)

(unit : logit)

Targeting for on-line adaptive recalibration will result in more difficult items being calibrated on responses of more able examinees, and easier items calibrated on responses of less able examinees. For those examinations in which subpopulations differ in their mean ability, this also means the confounding of calibration samples with group membership.

For example, if members of group A, on average, were more able than members of group B, more difficult items would tend to be given to group A, and easier items to group B. This would lead to the calibration of harder items using the responses of group A members, and easier items using the responses of group B members. If items functioned differentially for groups A and B, differential item functioning (DIF) would further compound the confounding of ability differences with group membership.

To simulate this confounding, **two subpopulations** were considered. They were:

(1) the reference group of examinees ("First-Time U.S."), and

(2) an ethnic group ("Ethnic Group") that is predominantly foreign-educated.

In the past, between 15% and 18% of the items in the examination have consistently demonstrated DIF against the Ethnic Group. Approximately 87% of the First-Time U.S. examinees in

this examination were whites, which is the reference group for DIF analyses for the examination.

The means and standard deviations of theta estimates for the Representative samples (for both the First-Time U.S. and Ethnic Group) are provided in Table 2. As expected, the Ethnic Group Representative samples have substantially different mean ability estimates than their First-Time U.S. counterparts. Overall (N = 2,100), the Ethnic Group Representative samples had a mean theta of -0.810, as opposed to the 0.071 mean for the Representative First-Time U.S. samples. Furthermore, the Ethnic Group Representative samples consistently had somewhat greater dispersion of theta estimates.

Four sample sizes were considered: 100, 200, 400, and 1,000. Except for N = 1,000 and whenever possible, three samples of the same size were obtained. In some cases, only two samples were produced due to insufficient case counts. Samples of 100, 200, and 400 were mutually exclusive. Samples of 1,000 were constructed by pooling samples of smaller sizes.

Analysis

Under the adaptive CAT conditions, the items in the simulated CAT forms were calibrated using responses from the corresponding CAT Samples. For instance, the items in the Hard CAT form were calibrated using responses from the High Ability CAT sample; the items in the Medium Difficulty CAT form were calibrated using the Middle Ability CAT Sample, and so forth. For comparison purposes, items in each of the CAT forms were also calibrated using the Representative Samples.

To simulate the PP and non-adaptive recalibration, parameter estimates were obtained for the items in the Shorter PP and Mini PP using responses from the Representative Samples. For purposes of comparison, they were also calibrated using two of the CAT Samples.

After responses were extracted for a given form from the source data, items in the form were calibrated using PARMATE to obtain a new set of b-values. The new b-values were then equated to the bank scale so that the mean of the new b-values would be equal to the mean of the corresponding bank b-values. Agreement between new b-values and bank b-values was assessed with two statistics: product-moment correlation (r) and the mean absolute difference (MAD). The MAD is the mean of absolute differences between the bank b-values and new b-values. Thus, if new b-values are very similar to bank b-values, the correlation typically will be higher and the MAD lower than if new b-values were dissimilar.

Results

Results for the First-Time U.S.: The results for the First-Time U.S. group are presented in Tables 3, 4, and 5, respectively, for the CAT forms, Mini PP, and Shorter PP. The Easy CAT form was not analyzed for the First-Time U.S. group because of insufficient cases in the ability range corresponding to the b-value range of easy items.

Two points must be born in mind in evaluating the results. First, as mentioned earlier, the mean ability estimate for the entire reference group who had taken the full-length PP form was 0.6. Second, the benchmark b-values also came from this group of First-Time U.S. educated examinees. Consequently, the results for the First-Time U.S. group will be better than those for the Ethnic Group.

Simulated CAT Forms: Table 3 shows that the correlations for the Hard form tended to be in the .80's and .90's, whereas those for the Medium Difficulty form tended to be between .50's and .70's. The MADs for the Hard form were in the .10's and those for the Medium Difficulty form were in the .10's and .20's. Thus, the Hard form produced b-values more similar to bank b-values.

The difference in correlations between the Hard versus Medium Difficulty forms would likely have been smaller if the Medium Difficulty form had contained items from as wide a range of difficulty as the items in the Hard form. Other things being equal, the less that range is restricted in one or both variables being correlated, the higher the correlation coefficient. The Hard form had twice the standard deviation of bank b-values for the Medium Difficulty form (.34 versus .17, as shown in Table 1).

The mean correlations for the Medium Difficulty form were predicted for the larger b-value standard deviation of .34, using the following formula:

$$r_{xy} = \frac{r'_{xy} \, (\sigma_x \, / \, \sigma'_x)}{\sqrt{1 - r'^{\,2}_{xy} + r'^{\,2}_{xy} \, (\sigma_x^2 \, / \, \sigma'^{\,2}_x)}} \, ,$$

where r'_{xy} is the restricted-range correlation, σ_x is the unrestricted-range standard deviation for the predictor, and σ'_x is the restricted-range standard deviation. The corrected correlations are presented in parentheses in the table. The correlations for the Hard form remained somewhat higher than the corrected correlations for the Medium Difficulty form. This was true for both the representative sample and the CAT sample.

The higher correlations and lower MADs for the Hard form may be attributed to two factors. First, the High Ability CAT Sample

in the theta range between -0.5 and 0.0 logit was more similar to the reference group (around 0.0) than was the Middle Ability CAT Sample between -1.0 and -0.5. Second, the items in the Hard form (b's between -0.63 and 0.86) enclosed the reference group mean ability, while the Medium Difficulty form did not (b's between -1.17 and -0.55). These findings seem to suggest that best calibration performance can be achieved for CAT forms if both the range of difficulties of items and the range of ability estimates of a sample include the mean ability of the reference group.

Within each form in Table 3, it is apparent that the Representative Samples from a wider ability range produced b-values that were more similar to the bank b-values than did the CAT Samples selected from a more narrow ability range. The differences in correlations between the Representative versus CAT samples were less notable with the Hard form than with the Medium Difficulty form. For example, with a sample size of 400, the average correlation was .933 for the Representative Sample and .915 for the High Ability CAT sample with the Hard form, while the corresponding values after correction were .931 and .827 with the Medium Difficulty form, respectively, for the Representative Sample and the Middle Ability CAT sample.

Considering the results for the Hard form, doubling the sample size compensated for the loss of accuracy caused by limiting the ability range of a sample. For example, the mean correlation between the Hard form and bank b-values for the Representative Sample of 100 examinees was .861. The corresponding correlation from the CAT sample of 100 examinees was somewhat lower (.807). With the CAT sample of 200, however, the mean correlation increased to a comparable value, .864. Similarly, the mean correlation for the CAT Sample of 400, .915, was higher than that for the Representative Sample of 200 (.895) but lower than the .933 correlation for the Representative Sample of 400. These results suggest that items similar in difficulty may be recalibrated using CAT responses if they are near the reference group mean ability and if a calibration sample is reasonably large.

In this research the three CAT forms were constructed simply by dividing the PP items into three groups. No <u>a priori</u> estimates were available as to what likely ranges of item difficulties for CAT tests would be. Although the CAT forms had a b-value standard deviation of .34, .17, and .23, there is no reason to believe that actual CAT forms will have as small (or large) dispersion as these forms. The correlations would be even greater if a CAT form had a greater standard deviation of b-values, relative to the standard deviation of b-values for the PP form.

Mini PP: Table 4 presents the results for the Mini PP. The correlations for the Mini PP were considerably higher than those in Table 3 for the simulated CAT forms. All the correlations for the Mini PP, except for one (.893), were in the .90's.

Since the Mini PP and the CAT forms contained similar numbers of items, test length cannot explain the better correlations for the Mini PP form. The better performance of the Mini PP is best explained by its substantially greater dispersion of b-values. Table 1 demonstrates that the bank b-values for the Mini PP had a standard deviation of .81, while the highest of the standard deviations for the three CAT forms was less than a half of the value (.34 for the Hard form).

Table 4 also reconfirmed the findings from the CAT forms. For a given sample size, the Representative Samples produced b-values more similar to bank b-values than did the CAT Samples. With the Mini PP, however, the differences tended to be very small between the Representative Samples and the CAT Samples in the High Ability range (between -0.5 and 0.0). With a sample size of 1,000, the difference in correlations between the High Ability CAT Sample and the Representative Sample was as small as .004 (.985 vs. .989). Thus, on-line adaptive recalibration is a more viable possibility with a form that contains items from varying difficulty levels than with a form with items similar in difficulty.

Shorter PP: The results for the Shorter PP, shown in Table 5, are almost identical to those for the Mini PP. This suggests that doubling test length from 75 items to 149 items did not improve the agreement between form b-values and bank b-values. Notice that the Mini PP and Shorter PP had almost identical means and standard deviations of bank b-values (-.97 vs. -.98 for the mean; .81 vs. .80 for SD, Table 1).

Results for the Ethnic Group: The results for the Ethnic Group are provided in Tables 6, 7, and 8 for the CAT forms, Mini PP, and Shorter PP, respectively. As reflected in the mean theta estimates of the Ethnic Group Representative Samples (Table 2), the performance level of the Ethnic Group has been substantially lower than that of the reference group of First-Time U.S. Moreover, as noted earlier, approximately 15% of items in the past PP examinations have been demonstrated to have DIF against the Ethnic Group (relative to whites who are the largest constituents of the First-Time U.S. group).

Simulated CAT Forms: As expected, the correlations and MADs indicated that the b-values for the CAT forms obtained using the Ethnic Group were substantially different from bank b-values.

The correlations for the Medium Difficulty and Easy forms were, for a large part, in the .10's, and the MADs in the .50's. The results for the Hard CAT form were somewhat better but still only moderate: correlations mostly in the .30's and the MADs mostly in the .40's. Once again, relative performance of the CAT forms seems to reflect the spread of b-values in the forms: the greater the spread, the greater the association between form b-values and bank b-values.

Comparisons of the results between the CAT Samples and Representative Samples within forms revealed that with the Medium Difficulty and Easy forms, the Representative Samples yielded b-values closer to bank b-values. However, with the Hard form, b-values from the CAT Samples were more similar to bank b-values. This is likely due to the Ethnic Group CAT Samples in the High Ability range having a mean ability that is more similar to that of the reference group than did the Ethnic Group Representative Samples. More specifically, as indicated in the footnotes of Table 6, the High Ability CAT Samples were selected from an ability range between -0.5 and 0.0. The reference group of First-Time U.S. has a mean in the vicinity of 0.0. The Ethnic Group Representative Samples had a mean between -.842 and -.731.

This implies that a sample that is more homogeneous in ability, but, on average, more similar to a reference group can produce b-values closer to the benchmark b-values than a sample that is more heterogenous but has a mean farther away from the reference group's mean. This finding seems to emphasize the importance of aligning the mean of a calibration sample with the mean of the reference group.

Mini PP: The correlations for the Mini PP, presented in Table 7, improved dramatically from those for the CAT forms in Table 6. Across different samples, the correlations for the Mini PP were primarily in the .60's, while those for the CAT forms were in the .30's at best. In contrast, the MADs exhibited only slight improvement. The differences between Table 6 and Table 7 reflect the effects of simultaneously calibrating items that are heterogeneous in difficulty, as opposed to homogeneous items.

As compared with the Representative Samples, b-values from the High Ability CAT Samples drawn from a theta band between -0.5 and 0.0 were, once again, more similar to bank b-values.

Shorter PP: The results for the Shorter PP are shown in Table 8. The correlations were roughly comparable to those for the Mini PP in Table 7. For the High Ability CAT Samples, the Shorter PP produced slightly higher correlations than did the Mini PP, irrespective of sample size. The MADs were consistently larger with the Shorter PP than with the Mini PP. These results once

again verified the earlier finding with the First-Time U.S. that increasing test length from 75 items to 149 items did not improve the correspondence between form b-values and bank b-values.

First-Time U.S. versus Ethnic Group: The results for the First-Time U.S. were compared with those from the Ethnic Group. Specifically, Table 3 was compared with Table 6, Table 4 with Table 7, and so forth. The differences in correlations between the groups for the CAT forms were substantial. The differences for the shorter versions of the PP, although much smaller, were still noticeably large. With the Mini and Shorter PP's, the correlations for the First-Time U.S. were in the .90's and those for the Ethnic Group were in the .60's and .70'. Those differences, and particularly the differences based on the CAT Samples, may likely be attributable to DIF against the Ethnic Group, because the same items were investigated using responses from examinees at the same ability level but from two different subpopulations.

Discussion and Conclusion

Recalibrating items using responses from CAT administrations (i.e., on-line adaptive recalibration) has potential as an alternative to (on-line) non-adaptive recalibration that uses responses from paper-and-pencil or linear-computerized administrations. CAT responses are inherently different from responses to paper-and-pencil items in that CAT responses would tend to come from only a portion of the total population of examinees (i.e., responses to difficult items from the able, responses to easy items from the less able). In comparison, paper-and-pencil responses will originate from examinees at various ability levels. The present study examined the effects of using responses from simulated CAT samples on Rasch b-values.

Overall, the results suggest problems with on-line adaptive recalibration as simulated in this study. Namely, the procedure involving recalibrating difficult items using CAT responses from able examinees, and separately calibrating easy items using CAT responses from less able examinees, seems to require some modifications. The results were reasonably good only when the mean of the item difficulties and the mean of examinee ability estimates were similar to the mean of the paper-and-pencil reference group.

There are two major reasons for the relatively poor results for the calibration of items using CAT responses. First, for some or even many items, CAT responses may come from examinees who are substantially different from a reference group -- different in terms of mean ability and group membership. Different group memberships may carry different amounts of DIF.

Second, for the "difficult-items-to-the-able: easy-items-to-the-less-able" on-line adaptive recalibration, only items of similar difficulty levels will be recalibrated simultaneously. In other words, difficult items will not be recalibrated with easy items, because the two sets of items will have been taken by two different groups of examinees. The resulting reduction in the dispersion of item difficulties will adversely impact parameter estimates.

It must be pointed out, however, that what was simulated in this study is rather "extreme," as compared with what will happen in actual CAT sessions. The simulated CAT forms in the present research were mutually exclusive, meaning that the forms had no items in common. In actual CAT sessions, however, it is often the case that examinees will be administered the same few initial items before branching out to different items. In a similar fashion, CAT sessions can be designed so that all examinees respond to a set of common items varying in difficulty (preferably scattered over a portion of a session to minimize the chance of security breach). An even better approach from a security's standpoint is the use of multiple sets of anchor items. If sets of common items are (quasi-)randomly administered to examinees, the groups will be randomly equivalent.

If each examinee takes a set of anchor items selected from a wide range of difficulty, it is possible to (re)calibrate all non-anchor items, both difficult and easy items, together, even though easy items have been taken only by less able examinees and difficult items by able examinees. It remains to be seen how much improvement will be achieved by such a simultaneous adaptive recalibration of all items using anchor items that vary in difficulty.

Thus, "modified" on-line adaptive recalibration may still be a possibility with the Rasch model. For modified on-line adaptive recalibration to be successful, the following four conditions seem essential. First, it is critical to define a reference group from which bank or benchmark item parameters have been drawn. Benchmark b-values for this study came from a large subpopulation taking the paper-and-pencil examination. Benchmark b-values substantially differed from the b-values obtained from another subpopulation for which relatively strong evidence of DIF exists.

Second, it is equally important to draw a CAT recalibration sample from the reference group. A CAT sample need <u>not</u> be representative of the reference group, that is, a CAT recalibration sample may have a smaller standard deviation of ability estimates than the reference group. However, the results seem to suggest that a CAT sample must have a mean ability comparable to that of the reference group. Specifically, a CAT sample that was more homogenous in ability and similar to the

reference group in terms of mean ability, produced b-values closer to bank b-values than did a paper-and-pencil random sample that was more heterogenous in ability but dissimilar to the reference group with regard to the mean theta.

Third, items to be (re)calibrated together must be relatively heterogenous in difficulty. The results improved substantially as the standard deviation of b-values for a set of items increased. The CAT forms simulated in this study had a b-value standard deviation about one-fifth to less than one half the standard deviation of the paper-and-pencil examination. Whether actual CAT forms will have a greater b-value standard deviation depends on various factors, such as stopping rules and the location of the initial item.

Items in a CAT form will typically be more similar in difficulty than items in a conventional form. However, by choosing an appropriate starting item (e.g., a relatively easy item for a relatively able reference group), many reference-group examinees will be administered a wider range of items.[1] Although this would mean a loss of efficiency, the amount of loss depends on several factors. If a large number of items can be recalibrated using responses straight from CAT sessions, the gain may be greater than the loss.[2]

Fourth, the results based on a CAT sample were either comparable or better than those based on a paper-and-pencil random sample of half the size, provided the first three conditions were satisfied. For instance, a CAT sample of 200 in the High Ability range yielded b-values more similar to benchmark b-values than a paper-and-pencil random sample of 100 (Table 4). With a sample size of 1,000, the differences between a CAT sample and a paper-and-pencil sample were negligible. Increasing the number of items in a form did not have a noticeable effect on b-values.

---

[1] This method will still leave out some items. These items may be administered to a calibration sample non-adaptively, along with a few other items of varying difficulties.

[2] The use of different starting points for different subpopulations may be another alternative. A lower starting point may be used for examinees from a reference group in order to obtain their responses to a set of diverse items. For examinees from other subpopulations whose responses will not be used for recalibration, an optimal starting point may be defined to maximize efficiency. However, the legal defensibility of this approach may be questioned, even if future research may demonstrate that starting points have no significant effects on examinee performance.

Since the present study was a live-data simulation, the findings need to be replicated with real CAT data. However, the results from this study indicate that "modified" on-line adaptive recalibration may be a viable and more economical method of item recalibration for CAT testing programs.

References

Lord, F.M. (1980). Applications of item responses theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. Psychometrika, 52, 589-618.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 298-325.

Table 1

Descriptive Statistics for the Forms Used

Form	# Items	Bank B-Value Mean	Bank B-Value SD	Bank B-Value Min	Bank B-Value Max	Abs. 1st Lds. Mean	Abs. 1st Lds. SD	Abs. 2nd Lds. Mean	Abs. 2nd Lds. SD
Simulated CAT Forms:									
Hard	73	-.02	.34	-.63	.86	.19	.09	.10	.08
Medium	73	-.82	.17	-1.17	-.55	.20	.10	.09	.07
Easy	68	-1.34	.23	-1.96	-.98	.19	.10	.10	.08
Mini PP	75	-0.98	.81	-3.03	.60	.21	.10	.10	.07
Shorter PP	149	-0.97	.80	-3.03	.63	.20	.10	.10	.08
PP	298	-0.97	.79	-3.06	.86	.20	.10	.10	.08

Table 2

Mean and SD of Thetas for Samples Representative of Subpopulations

	Ethnic Group		1st-Time, US	
N	Mean	SD	Mean	SD
400	-.834	.479	.050	.376
	-.826	.487	.091	.391
	-.783	.502	.049	.394
Mean	**-.814**	**.489**	**.063**	**.387**
200	-.831	.442	.102	.355
	-.782	.478	.061	.382
	-.842	.520	.084	.387
Mean	**-.818**	**.480**	**.082**	**.375**
100	-.786	.491	.102	.432
	-.807	.478	.042	.358
	-.731	.421	.104	.389
Mean	**-.775**	**.463**	**.083**	**.393**
2,100[1]	**-.810**	**.484**	**.071**	**.385**

[1] All cases combined. Because the smaller samples (i.e., N=400, 200, & 100) were drawn with no replacement, the combined total sample of 2,100 had no duplicate cases.

Table 3

Correlation B/w Bank b's and Part-form b's (r) & Their Mean Absolute Difference (MAD) :
Form = Simulated CAT Forms
Sample = First-time U.S.[1]

	Item Difficulty							
	Medium (-1.17 - -0.55)				Hard (-0.63 - 0.86)			
	Ability Range				Ability Range			
N	CAT Sample: Middle[2]		Repre-sentative[3]		CAT Sample: High[4]		Repre-sentative[3]	
	r	MAD	r	MAD	r	MAD	r	MAD
100	.486	.235	.537	.214	.819	.176	.833	.165
	.536	.252	.587	.205	.794	.190	.868	.162
			.539	.228			.882	.161
Mean	**.511**	**.244**	**.554**	**.216**	**.807**	**.183**	**.861**	**.163**
Corrected[5]	**(.765)**		**(.799)**					
200	.610	.230	.671	.157	.867	.152	.884	.146
	.586	.198	.547	.167	.861	.155	.887	.149
			.591	.140			.914	.134
Mean	**.598**	**.214**	**.603**	**.155**	**.864**	**.154**	**.895**	**.143**
Corrected[5]	**(.831)**		**(.834)**					
400	.643	.171	.779	.107	.921	.113	.933	.101
	.540	.196	.789	.113	.908	.123	.933	.096
			.795	.117			.932	.104
Mean	**.592**	**.184**	**.788**	**.112**	**.915**	**.118**	**.933**	**.100**
Corrected[5]	**(.827)**		**(.931)**					
1,000	**.622**	**.181**	**.828**	**.096**	**.934**	**.103**	**.953**	**.082**
Corrected[5]	**(.846)**		**(.947)**					

[1] The lowest theta range between -1.70 and -1.05 corresponding to the Easy CAT form contained only 141 First-Time U.S. examinees and was not analyzed.
[2] Between -1.0 and -0.5 logits.
[3] As shown in Table 2, mean thetas of the samples of various sizes ranged from .042 to .104. The standard deviations ranged from .355 to .432.
[4] Between -0.5 and 0.0 logits.
[5] Corrected for the smaller standard deviation of b-values, .17.

Table 4

Correlation B/w Bank b's and Part-form b's (r) & Their Mean Absolute Difference (MAD) :
Form = Mini PP
Sample = First-time U.S.

	Ability Range					
	CAT Sample: Middle[1]		CAT Sample: High[2]		Repre-sentative	
N	r	MAD	r	MAD	r	MAD
100	.915	.266	.944	.193	.951	.203
	.893	.302	.942	.215	.950	.212
					.930	.252
Mean	**.904**	**.284**	**.943**	**.204**	**.944**	**.222**
200	.920	.265	.961	.182	.978	.148
	.925	.246	.969	.154	.973	.164
					.977	.140
Mean	**.923**	**.256**	**.965**	**.168**	**.976**	**.151**
400	.935	.225	.979	.135	.981	.129
	.934	.239	.981	.129	.985	.114
					.984	.122
Mean	**.935**	**.232**	**.980**	**.132**	**.983**	**.122**
1000	.939	.228	.985	.119	.989	.097

[1] Between -1.0 and -0.5.
[2] Between -0.5 and 0.0.

Table 5

Correlation B/w Bank b's and Part-form b's (r) & Their Mean Absolute Difference (MAD) :
Form = Shorter PP
Sample = First-time U.S.

	Ability Range					
	CAT Sample: Middle[1]		CAT Sample: High[2]		Repre-sentative	
N	r	MAD	r	MAD	r	MAD
100	.915	.266	.944	.207	.933	.234
	.891	.295	.940	.218	.948	.223
					.939	.239
Mean	**.903**	**.281**	**.942**	**.213**	**.940**	**.232**
200	.926	.253	.962	.175	.968	.172
	.926	.238	.969	.151	.967	.174
					.974	.142
Mean	**.926**	**.246**	**.966**	**.163**	**.970**	**.163**
400	.938	.227	.979	.135	.983	.125
	.931	.241	.973	.143	.984	.112
					.982	.127
Mean	**.935**	**.234**	**.976**	**.139**	**.983**	**.121**
1000	**.941**	**.226**	**.983**	**.122**	**.989**	**.096**

[1] Between -1.0 and -0.5.
[2] Between -0.5 and 0.0.

Table 6

Correlation B/w Bank b's and Part-form b's (r) & Their Mean Absolute Difference (MAD) :
Form = Simulated CAT Forms
Sample = Ethnic Group

	Item Difficulty											
	Easy (-1.96 - -.98)				Mdm (-1.17 - -.55)				Hard (-.63 - .86)			
	Ability Range				Ability Range				Ability Range			
N	CAT Sample: Low[1]		Rep.[2]		CAT Sample: Middle[3]		Rep.[2]		CAT Sample: High[4]		Rep.[2]	
	r	MAD	r	MAD	r	MAD	r	MAD	r	MAD	r	MAD
100	.101	.581	.201	.552	.065	.582	.137	.477	.317	.494	.328	.491
	.125	.533	.224	.490	.138	.568	.108	.561	.342	.539	.296	.509
	.141	.553	.179	.530			.099	.537			.302	.487
Mean	**.122**	**.556**	**.201**	**.524**	**.102**	**.575**	**.115**	**.525**	**.330**	**.517**	**.309**	**.496**
200	.114	.562	.231	.500	.091	.525	.185	.507	.346	.468	.310	.512
	.112	.607	.273	.499	.101	.469	.106	.481	.367	.470	.334	.478
	.089	.559	.196	.554			.150	.502			.335	.498
Mean	**.105**	**.576**	**.233**	**.518**	**.096**	**.497**	**.147**	**.497**	**.357**	**.469**	**.326**	**.496**
400	.138	.561	.224	.524	.123	.516	.138	.461	.331	.505	.288	.491
	.130	.542	.208	.484	.116	.525	.185	.479	.364	.457	.348	.488
	.117	.579	.280	.489			.145	.477			.329	.509
Mean	**.128**	**.561**	**.237**	**.499**	**.120**	**.521**	**.156**	**.472**	**.348**	**.481**	**.322**	**.496**
1,000	**.131**	**.549**	**.222**	**.496**	**.114**	**.516**	**.169**	**.471**	**.350**	**.476**	**.321**	**.483**

[1] Between -1.7 and -1.05.
[2] Representative Samples. As shown in Table 2, mean thetas of the samples of various sizes ranged from -.842 to -.731. The standard deviations ranged from .421 to .520.
[3] Between -1.0 and -0.5.
[4] Between -0.5 and 0.0.

Table 7

Correlation B/w Bank b's and Part-form b's (r) &
Their Mean Absolute Difference (MAD) :
Form = Mini PP
Sample = Ethnic Group

	Ability Range					
	CAT Sample: Middle[1]		CAT Sample: High[2]		Repre-sentative	
N	r	MAD	r	MAD	r	MAD
100	.625	.570	.660	.427	.577	.462
	.652	.537	.695	.430	.669	.409
					.649	.404
Mean	**.639**	**.554**	**.678**	**.429**	**.632**	**.425**
200	.672	.535	.707	.389	.651	.425
	.674	.522	.691	.415	.662	.412
					.640	.455
Mean	**.673**	**.529**	**.699**	**.402**	**.651**	**.431**
400	.672	.532	.674	.439	.676	.389
	.650	.535	.700	.394	.664	.408
					.659	.405
Mean	**.661**	**.534**	**.687**	**.417**	**.666**	**.401**
1000	.668	.527	.695	.402	.671	.400

[1] Between -1.0 and -0.5.
[2] Between -0.5 and 0.0.

Table 8

Correlation B/w Bank b's and Part-form b's (r) & Their Mean Absolute Difference (MAD) :
Form = Shorter PP
Sample = Ethnic Group

	Ability Range					
	CAT Sample: Middle[1]		CAT Sample: High[2]		Repre-sentative	
N	r	MAD	r	MAD	r	MAD
100	.614	.603	.701	.514	.626	.560
	.616	.584	.708	.520	.667	.529
					.639	.544
Mean	**.615**	**.594**	**.705**	**.517**	**.644**	**.544**
200	.667	.532	.729	.479	.655	.547
	.644	.530	.712	.513	.671	.514
					.634	.549
Mean	**.656**	**.531**	**.721**	**.496**	**.653**	**.537**
400	.661	.540	.709	.509	.662	.515
	.659	.542	.729	.480	.662	.521
					.660	.527
Mean	**.660**	**.541**	**.719**	**.495**	**.661**	**.521**
1000	.666	.534	.726	.486	.666	.517

[1] Between -1.0 and -0.5.
[2] Between -0.5 and 0.0.

Abstract

Responses to previously calibrated items administered in a computerized adaptive testing (CAT) mode may be used to recalibrate the items. This live-data simulation study investigated the possibility, and limitations, of on-line adaptive recalibration of precalibrated items.

Responses to the items of a Rasch-based paper-and-pencil licensure examination were used to simulate CAT and paper-and-pencil administrations. To simulate CAT conditions, CAT forms having varying difficulty levels were defined, and CAT samples of varying sizes were selected from the ranges of ability that roughly corresponded to the difficulty levels of the items. To simulate paper-and-pencil test conditions, two shorter versions of the conventional examination were constructed, and representative samples of various sizes were drawn.

Responses to the CAT items were extracted from the CAT samples and, for comparison purposes, from the representative samples as well. Similarly, responses to the shorter paper-and-pencil forms were extracted from both the representative and CAT samples. Using those responses, the forms were calibrated and new b-values were compared with the bank b-values obtained from the responses of the reference group taking the paper-and-pencil examination.

The results indicate that bank b-values were not well replicated when difficult items were calibrated using responses from able examinees and easy items were calibrated using responses from less able examinees. On-line adaptive recalibration as simulated in this study has limitations.

However, a "modified" on-line adaptive recalibration may still be a possibility as long as (1) a reasonably large CAT recalibration sample is pre-defined from the reference group from which bank b-values have come; (2) the sample has a mean ability similar to that of the reference group, and (3) items to be recalibrated together are relatively heterogenous in Rasch difficulty. The third condition is counter to the notion that items in a CAT form will be targeted to ability and therefore relatively homogeneous in difficulty. Ways to simultaneously recalibrate items from a wider range of difficulty using CAT responses are discussed.